AD703002

FTD-MT-24-409-69

AN EXPERIMENT OF CREATION OF AN INFORMATION RETRIEVAL LANGUAGE ON COMPUTER ENGINEERING

by

V. K. Vakhabov, A. A. Mikhaylova, et al.

D D C
APR 2 1970
A

Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce, for sale to the general public.

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151

3

FTD-MT-24-409-69

EDITED MACHINE TRANSLATION

AN EXPERIMENT OF CREATION OF AN INFORMATION RETRIEVAL LANGUAGE ON COMPUTER ENGINEERING

By: V. K. Vakhabov, A. A. Mikhaylova, et al.

English pages: 9

Source: Vsesoyuznaya Konferentsiya po Informatsionno-Poiskovym Sisteman i Avtomatizirovannoy Obrabotke Nauchno-Teknicheskoy Informatsii, 3d, Moscow, 1966, Trudy (Transactions of the Third All-Union Conference on Information Retrieval Systems and Automatic Processing of Scientific and Technical Information. Moscow, 1966). 1967, pp. 156-161.

This document is a Mark II machine aided translation, post-edited for technical accuracy by: W. W. Kennedy

THIS TRANSLATION IS A RENDITION OF THE ORIGINAL FOREIGN TEXT WITHOUT ANY ANALYTICAL OR EDITORIAL COMMENT. STATEMENTS OR THEORIES ADVOCATED OR IMPLIED ARE THOSE OF THE SOURCE AND DO NOT NECESSARILY REFLECT THE POSITION OR OPINION OF THE FOREIGN TECHNOLOGY DIVISION.	PREPARED BY: TRANSLATION DIVISION FOREIGN TECHNOLOGY DIVISION WP-AFB, OHIO.

FTD-MT- 24-409-69 Date 31 Dec 19 69

U. S. BOARD ON GEOGRAPHIC NAMES TRANSLITERATION SYSTEM

Block	Italic	Transliteration	Block	Italic	Transliteration
А а	*А а*	A, a	Р р	*Р р*	R, r
Б б	*Б б*	B, b	С с	*С с*	S, s
В в	*В в*	V, v	Т т	*Т т*	T, t
Г г	*Г г*	G, g	У у	*У у*	U, u
Д д	*Д д*	D, d	Ф ф	*Ф ф*	F, f
Е е	*Е е*	Ye, ye; E, e*	Х х	*Х х*	Kh, kh
Ж ж	*Ж ж*	Zh, zh	Ц ц	*Ц ц*	Ts, ts
З з	*З з*	Z, z	Ч ч	*Ч ч*	Ch, ch
И и	*И и*	I, i	Ш ш	*Ш ш*	Sh, sh
Й й	*Й й*	Y, y	Щ щ	*Щ щ*	Shch, shch
К к	*К к*	K, k	Ъ ъ	*Ъ ъ*	"
Л л	*Л л*	L, l	Ы ы	*Ы ы*	Y, y
М м	*М м*	M, m	Ь ь	*Ь ь*	'
Н н	*Н н*	N, n	Э э	*Э э*	E, e
О о	*О о*	O, o	Ю ю	*Ю ю*	Yu, yu
П п	*П п*	P, p	Я я	*Я я*	Ya, ya

* ye initially, after vowels, and after ъ, ь; e elsewhere. When written as ё in Russian, transliterate as yё or ё. The use of diacritical marks is preferred, but such marks may be omitted when expediency dictates.

AN EXPERIMENT OF CREATION OF AN INFORMATION RETRIEVAL LANGUAGE ON COMPUTER ENGINEERING

V. K. Vakhabov, A. A. Mikhaylova, L. M. Yesilevskaya, and T. S. Kutayeva

At the Perm Scientific Research Institute of Control Machines and Systems is being conducted work on creation of an automated information-retrieval [IPS] (ИПС) for a reference and information fund (SIF) of the [ONTI] (ОНТИ) [Joint Scientific and Technical Publishing House] of instruments.

An important element of the IPS is its information retrieval language [IPYa] (ИРЯ). In the present report is reported the development of information retrieval language of the descriptor type for the "Computer engineering" subject field.

In the selection of IPYa structure were taken into account the following features of IPS operation:

1. High productivity of search (up to three-four thousand inquiries per day) through the application of a magnetic drum electronic digital computer.

2. The need for machine translation of key words of the search instruction into descriptor codes to increase search productivity.

3. Absence of direct feedback with the consumer during machine search for correction of search instruction for the purpose of obtaining required completeness and accuracy. Feedback is attained in the system through the application of a three-circuit IPS system, where the first circuit is not embraced by feedback (Fig. 1). Under such conditions the presence of noise requires only certain increase of the productivity of the secondary circuit. Therefore, the value of noise is a less important characteristic than the value of losses.

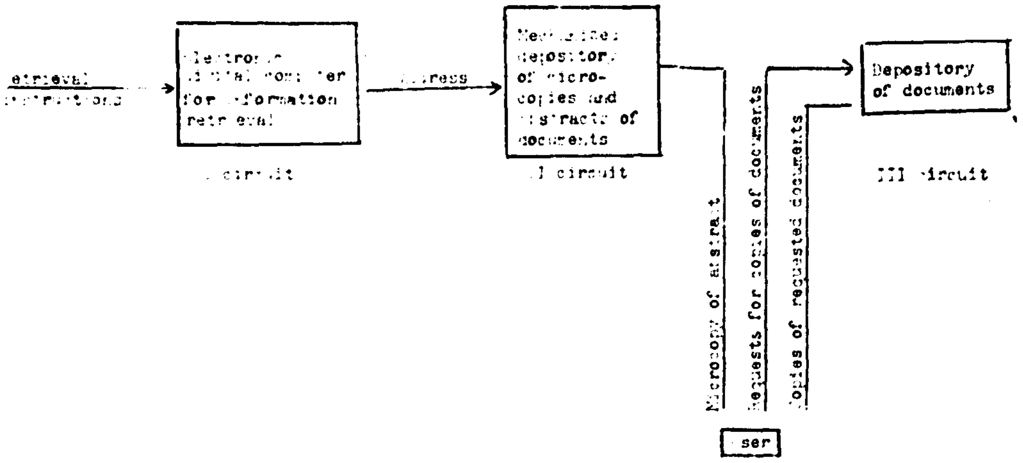
Retrieval instructions
Electronic digital computer for information retrieval
Address
Mechanized depository of microcopies and abstracts of documents
Depository of documents
I circuit
II circuit
III circuit
Microcopy of abstract
Requests for copies of documents
Copies of requested documents
User

Fig. 1. The block diagram of a three-circuit IPS.

4. The whole information array in the branch center is broken up into a number of big thematic subfiles with a volume of the order of 30 thousand documents. For each subfile is developed a local IPYa.

5. Presence of a machine dictionary for translation of key words of an inquiry into codes of descriptors inevitably requires certain standardization of key words of dictionary and search instruction.

The following basic rules were adopted:

a) the overwhelming majority of key words used are individual words of natural language. A word combination is used only when it is a commonly used scientific term. It can correspond to an abbreviation, which is also included in the dictionary. For example: computer - VM, memory unit - ZU;

b) key words have to be nouns, adjectives, and rarely numerals;

c) all words are used in the singular with the exeption of those words which do not have a singular;

d) adjectives are used in the masculine gender. The dictionary was composed on a file of 1300 abstracts on computing machinery from RET [Translator's Note: Expansion unknown] journals.

In the process of free indexing of abstracts key words were selected, they were unified into classes of conditional equivalence, and basic connections in the form of references "see" to higher descriptors were established. After termination of this work the dictionary contained 664 words and 367 descriptors (classes of conditional equivalence). Then on the basis of available dictionary were indexed 1060 more abstracts. The dictionary was supplemented with new words. At present the dictionary contains 702 key words united into 404 classes of conditional equivalence.

From these data it is possible to trace the character of dependence of the value of the dictionary on the volume of the information file (Fig. 2). From the given graph it is clear that growth of dictionary with increase of the file of documents is considerably delayed. This phenomenon is called dictionary saturation.

To evaluate the developed language an experiment was tried on an initial file of 1300 documents and on a total of 2360 documents. The purpose of the experiment was:

1) to determine the value of noise factors and losses;

2) to clarify how these indices vary with growth of the search file. This is necessary to forecast the input of grammatical means into the IPYa.

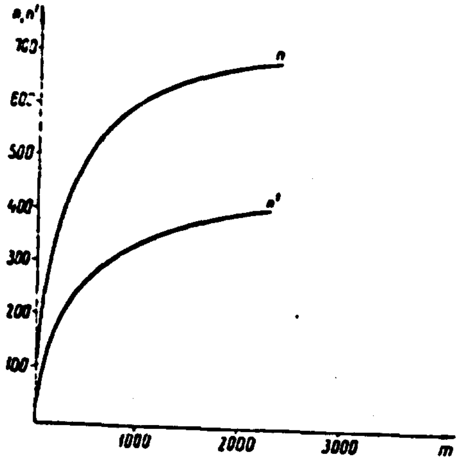

Fig. 2. Dependence of volume of descriptor dictionary and dictionary of key words on the number of documents in the file: m is the number of documents in the file. n' is the number of descriptors in the dictionary, n is the number of key words in the dictionary.

For the experiment were formulated 250 inquiries. They were composed by specialists not participating in IPYa development. In every inquiry were at least three key words, a maximum of nine, and an average of five. On the basis of these inquiries the coefficient of accuracy was calculated by the formula [2]:

$$A = \frac{100R}{L}\%,$$

where R is the number of relevant documents in delivery; L is the total number of documents in a delivery.

To calculate the coefficient of accuracy was performed a search on 150 inquiries. On all 150 inquiries fell 412 documents, 351 of them relevant.

Analysis showed that on 113 demands fell only relevant documents, and on 37 inquiries, besides the relevant one documents not responding to an inquiry.

Out of 37 inquiries on 22 fell one unnecessary document, on 10 two, and on the other five inquiries 3 or more documents. After study of causes of errors it turned out that only 7% of information noise was caused by indexing deficiencies, and 93% is irremovable noise due to false combinations.

Example 1. Inquiry: the principle of action of ferrite-core ZU. The total number of documents issued on the inquiry is six, five relevant. One document does not correspond to the inquiry and was issued as a result of false combinations. In the document is discussed the principle of action of thin-film ZU and the method of selection of words with the help of ferrite cores.

Example 2. Inquiry: characteristics of magnetic memory units. The total number of documents issued on the inquiry is 22, 21 relevant; one superfluous (not responding to the inquiry) document is issued because of false combinations. In the document are discussed the characteristics of an electronic miniature system of military assignment with magnetic ZU.

The coefficient of completeness was calculated by two methods: first, as a percentage of the number of relevant documents in the delivery to the total number of relevant documents in the search file [2]; second as a ratio of the number of found initial documents to 100 inquiries [2]. The initial document is considered the document by which is composed the inquiry. There are 100 initial documents.

In view of the complexity and time-consuming nature of finding the total number of relevant documents in the search file, the coefficient of completeness was determined by the first method for 10 inquiries. Coefficients of completeness calculated by two different methods give the same result (see table).

Losses of documents occur as a result of the complexity of taking into account all aspects illuminated in the document during indexing.

Parameters of IPS effectiveness	Information file on which IPYa was created (1300), %	Total file on which IPYa was worked (2360), %
Coefficient of completeness (first method)	92	92
Coefficient of completeness (second method)	92	92
Coefficient of accuracy	85	80

Example. Inquiry: use of thermoregulators for normal ZU operation. Due to the absence in the search image of the key word "Thermoregulator" in answer to this inquiry was not issued the following test:

Copy No. 9967. (Omission here) Universal Decimal Classification 681.142.652.2. Circuit stabilization of recording current of an address circuit of a ZU of the Z type. "Information reference sheets." 1963 No. 3441, 3 p., illustrated.

A description is given of a circuit for stabilization of address recording current intended for use in ZU of the Z type with 128 27-digit numbers consisting of ferrite cores of the VT-1 type (2 × 1, 4 × 0.9). The cores work in the following conditions: readout current $I_{\text{сч}} = -(1.2\text{-}1.5)$ A; the digit recording current $I_{\text{разр}} = 0.6$ A; the address recording current $I_{\text{зап}} = 0.7$ A; fixed bias $I_{\text{см}} = -0.3$ A. The stabilization circuit consists of six P25B (or P25A) transistors and regardless of the number of reversed cores in the numerical rule (of the value of load) ensures stable value of current $I_{\text{зап}} = 0.7$ A by limitation of it by the internal resistance of the circuit. To ensure normal ZU operation during change of ambient temperature in a stabilizing circuit are used thermoregulators with two semiconductor thermistors and a diode. The proposed circuit provides normal ZU operation in the range of temperatures from -20 to +60°C. The results of the experiment are given in the table.

In the process of indexing the information file was studied the law of distribution of descriptors in the search images of documents.

As is known, [5] and [6], the frequency of appearance of words of natural language with high accuracy follows the Zipf law:

$$P_i = \frac{K}{i},$$

where K = const; i = 1, 2, ..., n is the ordinal number of the word with location in order of decreasing frequency.

More exactly the distribution of words of natural language is described by the Mandel'brot law, which essentially generalizes the Zipf law:

$$P_i = \frac{K}{(B+i)^{\alpha}},$$

where K, B and α are constant, where $1 \leq \alpha \leq 1.2$.

Study of the real law of distribution of descriptors in search images of documents shows that the Zipf and Mandel'brot laws, well-known for distribution of the words of natural language, well describe also distribution of descriptors.

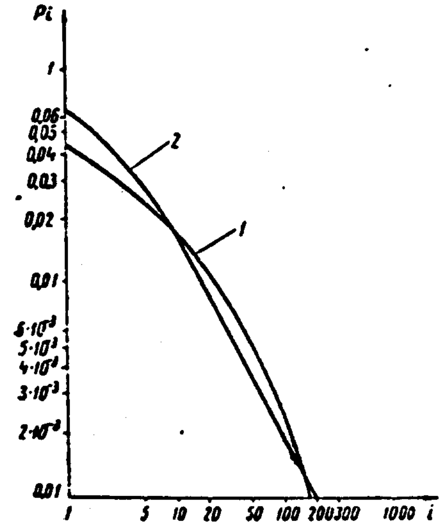

Fig. 3. 1 – real law of distribution of descriptors in search images of documents, 2 – curve plotted according to the law $P_i = \frac{0.202}{2 + i}$.

On Fig. 3 is shown the real law of distribution of descriptors. From the graph it is clear that this law can be dercribed by the expression:

$$P_i \approx \frac{0.202}{2+i}.$$

which will be in complete agreement with the Zipf and Mandel'brot laws.

Conclusions

1. Conducted experiments show that the developed IPYa, in spite of simplicity of structure, has fully satisfactory characteristics.

2. Growth of dictionary is considerably delayed when the file goes over 2000 documents.

3. Input of grammatical means into the IPYa for small files is inexpedient. The question of the need for introduction of grammatical means into the developed IPYa will be examined after the carrying out in 1967 of experiments on a file of 15-20 thousand documents.

4. Distribution of descriptors in search images of documents obeys the Zipf and Mandel'brot laws as word in natural language texts.

Bibliography

1. Vickery B., Vocabularies for coordinate sistems. «Aslib. Proc.», 1963, 15, № 6, 170—176
2. Cleverdon C., Lancaster F., Mills I. Uncovering some facts of life in information retrieval «Spec. Libr», 1964, 55, № 2, 86—91.
3. Lancaster F. W. Same observations on the performance of EJC role indikators in a mechanized retrieval sistem. «Spec. Libr.», 1964, 55, № 10, 696—701

4. Mikhaylov A. I., Chernyy A. I., Gilyarevskiy R. S. "Osnovy nauchnoy informatsii" ("Bases of scientific information"). M., Izd. "Nauka", 1965.

5. Brillyuyen L. "Nauka i teoriya informatsii" (Science and information theory"). M., Fizmatgiz, 1960.

6. Mandel'brot B. O rekurrentnom kodirovanii, ogranichivayushchem vliyaniye pomekh (About recurrence coding limiting interference effect). In the collection. "Teoriya peredachi soobshcheniy". M. IL., 1957.

DATA HANDLING PAGE

01-ACCESSION NO. TT0500	98-DOCUMENT LOC	39-TOPIC TAGS information storage and retrieval, computer design, magnetic drum, computer language
09-TITLE AN EXPERIMENT OF CREATION OF AN INFORMATION RETRIEVAL LANGUAGE ON.COMPUTER ENGINEERING -U-		
47-SUBJECT AREA 09, 14		

42-AUTHOR CO-AUTHORS VAKHAROV, V. K. ; 16-MIKHAYLOVA, A. A. ; 16-YESILEVSKAYA, L. M. ; 16-KUTAYEVA, T. S.	10-DATE OF INFO -----67
43-SOURCE VSESOYUZNAYA KONFERENTSIYA PO INFORMATSIONNO-POISKOVYM SISTEMAM I AVTOMATIZIROVANNOY OBRABOTKE NAUCHNO-TEKNICHESKOY INFORMATSII, 3D MOSCOW, 1966, TRUDY (RUSSIAN)	68-DOCUMENT NO. FTD-MT-24-409-69 69-PROJECT NO. 6050205

63-SECURITY AND DOWNGRADING INFORMATION UNCL. 0	64-CONTROL MARKINGS NONE	97-HEADER CLASN UNCL

76-REEL/FRAME NO. 1891 0457	77-SUPERSEDES	78-CHANGES	40-GEOGRAPHICAL AREA UR	NO. OF PAGES 9
CONTRACT NO.	X REF ACC. NO. 65-AT8020963	PUBLISHING DATE 94-	TYPE PRODUCT TRANSLATION	REVISION FREQ NONE

STEP NO.
02-UR/0000/67/000/000/0156/0161

ABSTRACT

(U) At the Perm Scientific Research Institute of Control Mechanisms and Systems, an information retrieval language of the descriptor type was designed for the subject field "computer engineering" to exploit a magnetic drum digital computer for 3-4,000 inquiries per day with computer translation of the keywords of the retrieval image into descriptor codes in the absence of corrective feedback from the user during the search. The system would process a number of subject subcorpuses, each comprising about 30,000 documents, with a local information retrieval language for each. A hierarchical organization of concepts and the introduction of grammatical resources were intended to reduce false drops. The initial experiment with 2,360 documents indicated no advantage of grammar, but further experiments to clarify this point are projected to comprise 15-20,000 documents. The descriptor image of each document has an average of 8-10 descriptors. The computer dictionary for translating keywords into descriptor codes consists mainly of nouns and adjectives of the natural language or their abbreviations. The dictionary was initiated in a conventional indexing stage covering 1,300 abstracts, which produced 664 words and 367 descriptions. Orig. art. has: 3 figures, 1 table, and 4 formulas.